第４講　データの分析

[1] (1) 次の [ア]，[イ] に当てはまるものを，下の⓪～⑤のうちから一つずつ選べ。ただし，解答の順序は問わない。

99 個の観測値からなるデータがある。四分位数について述べた記述で，どのようなデータでも成り立つものは [ア] と [イ] である。

⓪ 平均値は第 1 四分位数と第 3 四分位数の間にある。

① 四分位範囲は標準偏差より大きい。

② 中央値より小さい観測値の個数は 49 個である。

③ 最大値に等しい観測値を 1 個削除しても第 1 四分位数は変わらない。

④ 第 1 四分位数より小さい観測値と，第 3 四分位数より大きい観測値とをすべて削除すると，残りの観測値の個数は 51 個である。

⑤ 第 1 四分位数より小さい観測値と，第 3 四分位数より大きい観測値とをすべて削除すると，残りの観測値からなるデータの範囲はもとのデータの四分位範囲に等しい。

（[1]は次々ページに続く。）

< NOTE >

(2)　図１は，平成 27 年の男の市区町村別平均寿命のデータを 47 の都道府県 P1，P2，…，P47 ごとに箱ひげ図にして，並べたものである。

次の(Ⅰ)，(Ⅱ)，(Ⅲ)は図１に関する記述である。

(Ⅰ)　四分位範囲はどの都道府県においても１以下である。

(Ⅱ)　箱ひげ図は中央値が小さい値から大きい値の順に上から下へ並んでいる。

(Ⅲ)　P1 のデータのどの値と P47 のデータのどの値とを比較しても 1.5 以上の差がある。

次の［ウ］に当てはまるものを，下の⓪～⑦のうちから一つ選べ。

(Ⅰ)，(Ⅱ)，(Ⅲ)の正誤の組合せとして正しいものは［ウ］である。

	⓪	①	②	③	④	⑤	⑥	⑦
(Ⅰ)	正	正	正	誤	正	誤	誤	誤
(Ⅱ)	正	正	誤	正	誤	正	誤	誤
(Ⅲ)	正	誤	正	正	誤	誤	正	誤

(１は次ページに続く。)

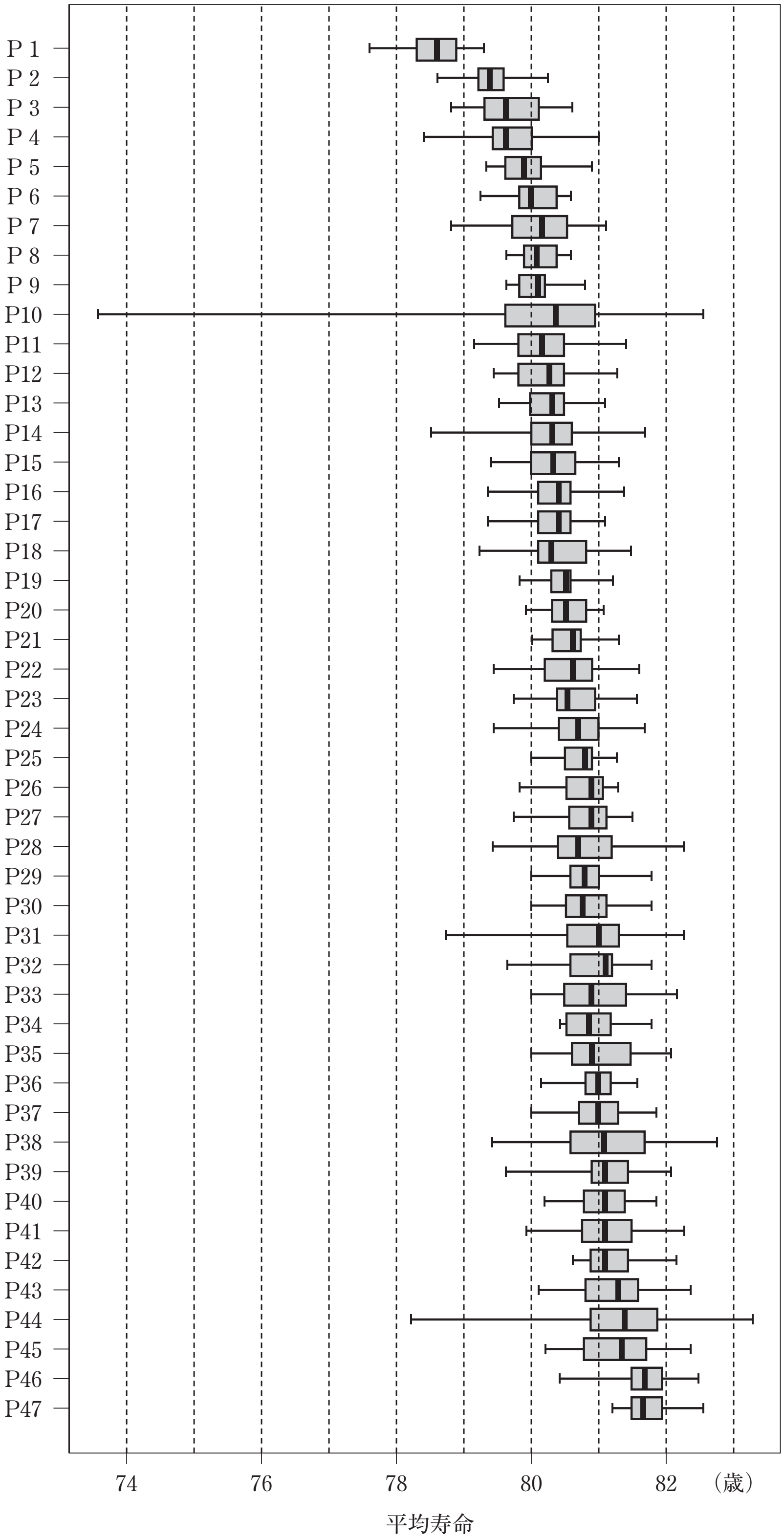


図１　男の市区町村別平均寿命の箱ひげ図

（出典：厚生労働省の Web ページにより作成）

（１は次ページに続く。）

(3)　ある県は 20 の市区町村からなる。図 2 はその県の男の市区町村別平均寿命のヒストグラムである。なお，ヒストグラムの各階級の区間は，左側の数値を含み，右側の数値を含まない。

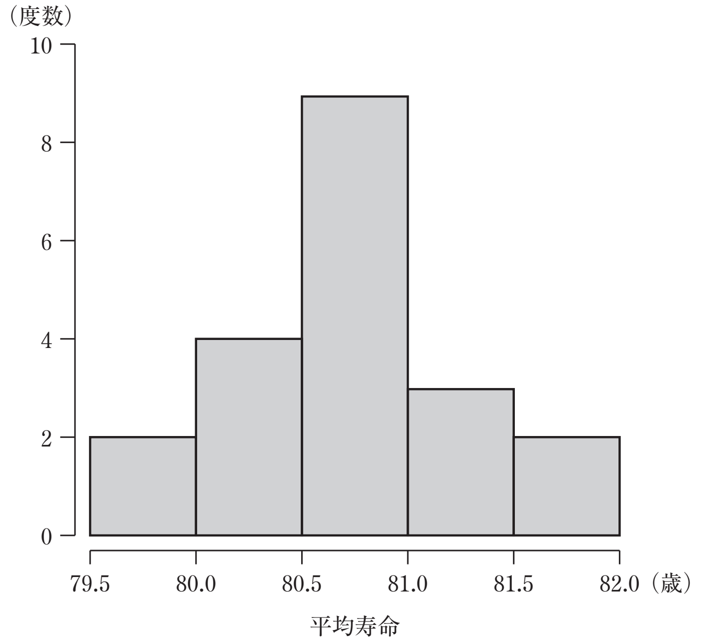
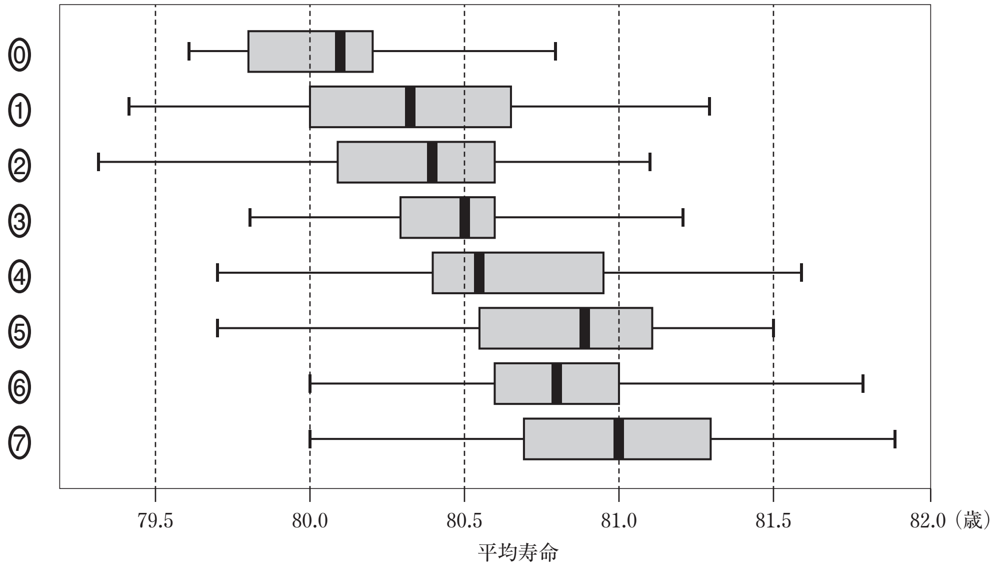


図 2　市区町村別平均寿命のヒストグラム
(出典：厚生労働省の Web ページにより作成)

(1は次ページに続く。)

次の［エ］に当てはまるものを，下の⓪～⑦のうちから一つ選べ。

図２のヒストグラムに対する箱ひげ図は［エ］である。



（1は次ページに続く。）

(4) 図 3 は，平成 27 年の男の都道府県別平均寿命と女の都道府県別平均寿命の散布図である。2 個の点が重なって区別できない所は黒丸にしている。図には補助的に切片が 5.5 から 7.5 まで 0.5 刻みで傾き 1 の直線を 5 本付加している。

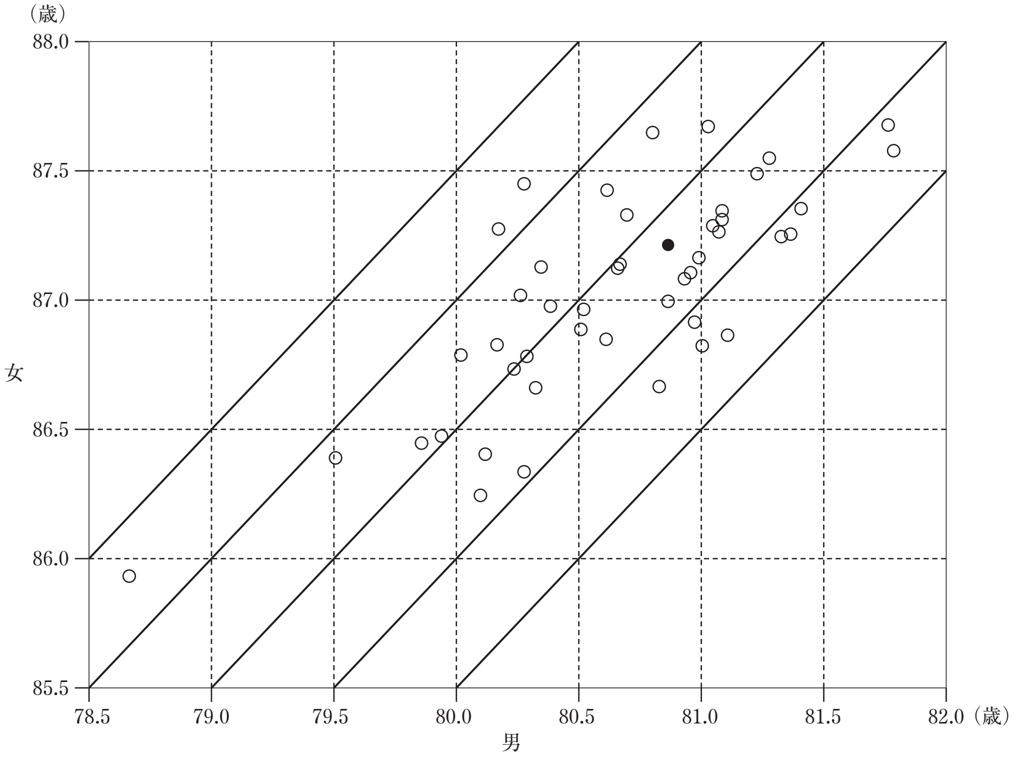


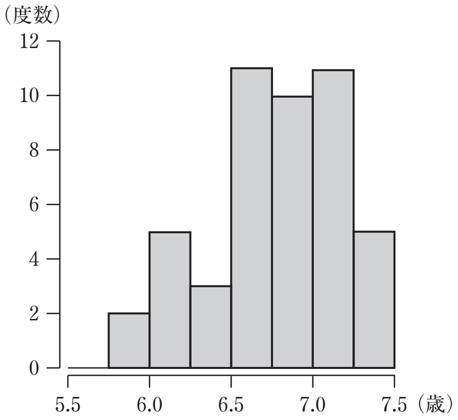
図 3　男と女の都道府県別平均寿命の散布図
(出典：厚生労働省の Web ページにより作成)

(1は次ページに続く。)

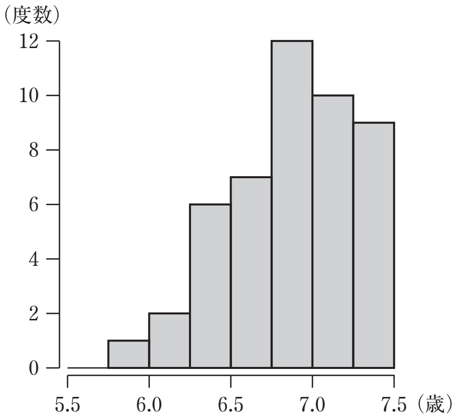
次の オ に当てはまるものを，下の⓪～③のうちから一つ選べ。

都道府県ごとに男女の平均寿命の差をとったデータに対するヒストグラムは オ である。なお，ヒストグラムの各階級の区間は，左側の数値を含み，右側の数値を含まない。

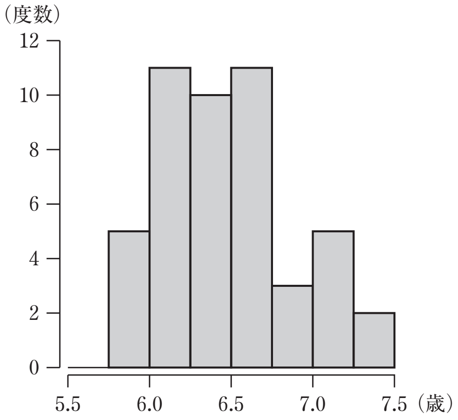
⓪



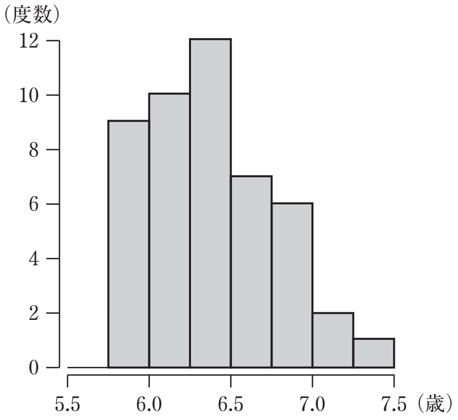
①



②



③



2　太郎さんと花子さんは二つの変量 x，y の相関係数について考えている。二人の会話を読み，下の問いに答えよ。

> 花子：先生からもらった表計算ソフトのＡ列とＢ列に値を入れると，Ｅ列にはＤ列に対応する正しい値が表示されるよ。
>
> 太郎：最初は簡単なところで二組から考えてみよう。
>
> 花子：2 行目を $(x,\ y)=(1,\ 2)$，3 行目を $(x,\ y)=(2,\ 1)$ としてみるね。

このときのコンピュータの画面の様子が次の図である。

	A	B	C	D	E
1	変量 x	変量 y		(x の平均値)＝	ア
2	1	2		(x の標準偏差)＝	イ
3	2	1		(y の平均値)＝	ア
4				(y の標準偏差)＝	イ
5					
6				(x と y の相関係数)＝	ウ
7					

(1)　ア，イ，ウ に当てはまるものを，次の⓪～⑨のうちから一つずつ選べ。ただし，同じものを繰り返し選んでもよい。

⓪　−1.50　①　−1.00　②　−0.50　③　−0.25　④　0.00
⑤　0.25　⑥　0.50　⑦　1.00　⑧　1.50　⑨　2.00

（2は次ページに続く。）

太郎：3 行目の変量 y の値を 0 や −1 に変えても相関係数の値は［ウ］になったね。 花子：今度は，3 行目の変量 y の値を 2 に変えてみよう。 太郎：エラーが表示されて，相関係数は計算できないみたいだ。

(2)　変量 x と変量 y の値の組を変更して，$(x,\ y)=(1,\ 2),\ (2,\ 2)$ としたときには相関係数が計算できなかった。その理由として最も適当なものを，次の⓪～③のうちから一つ選べ。［エ］

⓪　値の組の個数が 2 個しかないから。
①　変量 x の平均値と変量 y の平均値が異なるから。
②　変量 x の標準偏差の値と変量 y の標準偏差の値が異なるから。
③　変量 y の標準偏差の値が 0 であるから。

花子：3 行目の変量 y の値を 3 に変更してみよう。相関係数の値は 1.00 だね。 太郎：3 行目の変量 y の値が 4 のときも 5 のときも，相関係数の値は 1.00 だ。 花子：相関係数の値が 1.00 になるのはどんな特徴があるときかな。 太郎：値の組の個数を多くすると何かわかるかもしれないよ。

（②は次ページに続く。）

花子：じゃあ，次の値の組の個数を 3 としてみよう。
太郎：$(x,\ y)=(1,\ 1),\ (2,\ 2),\ (3,\ 3)$ とすると相関係数の値は 1.00 だ。
花子：$(x,\ y)=(1,\ 1),\ (2,\ 2),\ (3,\ 1)$ とすると相関係数の値は 0.00 になった。
太郎：$(x,\ y)=(1,\ 1),\ (2,\ 2),\ (2,\ 2)$ とすると相関係数の値は 1.00 だね。
花子：まったく同じ値の組が含まれていても相関係数の値は計算できることがあるんだね。
太郎：思い切って，値の組の個数を 100 にして，1 個だけ $(x,\ y)=(1,\ 1)$ で，99 個は $(x,\ y)=(2,\ 2)$ としてみるね……。相関係数の値は 1.00 になったよ。
花子：値の組の個数が多くても，相関係数の値が1.00になるときもあるね。

(3)　相関係数の値についての記述として**誤っているもの**を，次の⓪～④のうちから一つ選べ。 **オ**

⓪　値の組の個数が 2 のときには相関係数の値が 0.00 になることはない。
①　値の組の個数が 3 のときには相関係数の値が -1.00 となることがある。
②　値の組の個数が 4 のときには相関係数の値が 1.00 になることはない。
③　値の組の個数が 50 であり，1 個の値の組が $(x,\ y)=(1,\ 1)$，残りの 49 個の値の組が $(x,\ y)=(2,\ 0)$ のときは相関係数の値は -1.00 である。
④　値の組の個数が 100 であり，50 個の値の組が $(x,\ y)=(1,\ 1)$，残りの 50 個の値の組が $(x,\ y)=(2,\ 2)$ のときは相関係数の値は 1.00 である。

(2は次ページに続く。)

花子：値の組の個数が 2 のときは，相関係数の値は 1.00 か [ウ]，または計算できない場合の 3 通りしかないね。

太郎：値の組を散布図に表したとき，相関係数の値はあくまで散布図の点が [カ] 程度を表していて，値の組の個数が 2 の場合に，花子さんが言った 3 通りに限られるのは [キ] からだね。値の組の個数が多くても値の組が 2 種類のときはそれらにしかならないんだね。

花子：なるほどね。相関係数は，そもそも値の組の個数が多いときに使われるものだから，組の個数が極端に少ないときなどにはあまり意味がないのかもしれないね。

太郎：値の組の個数が少ないときはもちろんのことだけど，基本的に散布図と相関係数を合わせてデータの特徴を考えるとよさそうだね。

(4) [カ]，[キ] に当てはまる最も適当なものを，次の各解答群のうちから一つずつ選べ。

[カ] の解答群

⓪ x 軸に関して対称に分布する

① 変量 x，y のそれぞれの中央値を表す点の近くに分布する

② 変量 x，y のそれぞれの平均値を表す点の近くに分布する

③ 円周に沿って分布する

④ 直線に沿って分布する

[キ] の解答群

⓪ 変量 x の中央値と平均値が一致する

① 変量 x の四分位数を考えることができない

② 変量 x，y のそれぞれの平均値を表す点からの距離が等しい

③ 平面上の異なる 2 点は必ずある直線上にある

④ 平面上の異なる 2 点を通る円はただ 1 つに決まらない

3　地方の経済活性化のため，太郎さんと花子さんは観光客の消費に着目し，その拡大に向けて基礎的な情報を整理することにした。以下は，都道府県別の統計データを集め，分析しているときの二人の会話である。会話を読んで下の問いに答えよ。ただし，東京都，大阪府，福井県の 3 都府県のデータは含まれていない。また，以後の問題文では「道府県」を単に「県」として表記する。

> 太郎：各県を訪れた観光客数を x 軸，消費総額を y 軸にとり，散布図を作ると図 1 のようになったよ。
> 花子：消費総額を観光客数で割った消費額単価が最も高いのはどこかな。
> 太郎：元のデータを使って県ごとに割り算をすれば分かるよ。
> 　　　北海道は……。44 回も計算するのは大変だし，間違えそうだな。
> 花子：図 1 を使えばすぐ分かるよ。

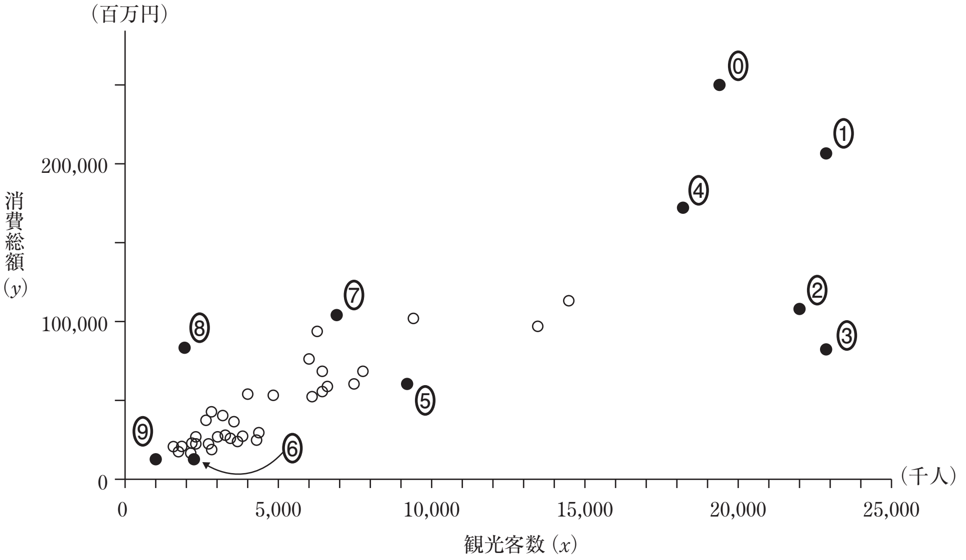


図 1

（3は次ページに続く。）

(1)　図１の観光客数と消費総額の間の相関係数に最も近い値を，次の⓪～④のうち から一つ選べ。 ア

⓪　-0.85　①　-0.52　②　0.02　③　0.34　④　0.83

(2)　44 県それぞれの消費額単価を計算しなくても，図１の散布図から消費額単価が最も高い県を表す点を特定することができる。その方法を，「直線」という単語を用いて説明せよ。

(3)　消費額単価が最も高い県を表す点を，図１の⓪～⑨のうちから一つ選べ。
イ

(3は次ページに続く。)

花子：元のデータを見ると消費額単価が最も高いのは沖縄県だね。沖縄県の消費額単価が高いのは，県外からの観光客数の影響かな。

太郎：県内からの観光客と県外からの観光客とに分けて 44 県の観光客数と消費総額を箱ひげ図で表すと図 2 のようになったよ。

花子：私は県内と県外からの観光客の消費額単価をそれぞれ横軸と縦軸にとって図 3 の散布図をつくってみたよ。沖縄県は県内，県外ともに観光客の消費額単価は高いね。それに，北海道，鹿児島県，沖縄県は全体の傾向から外れているみたい。

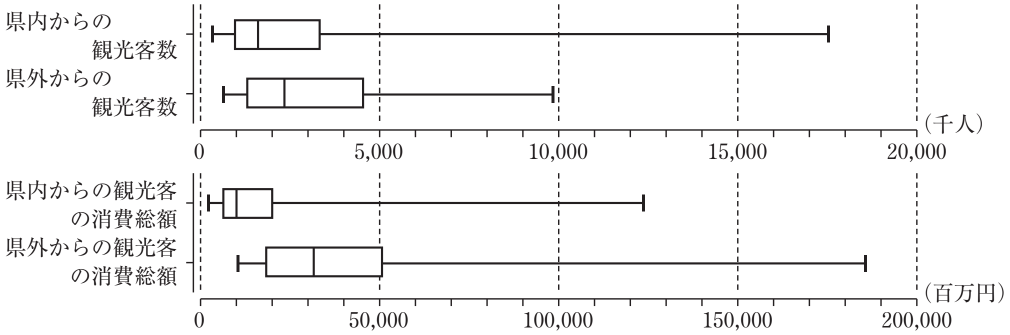


図 2

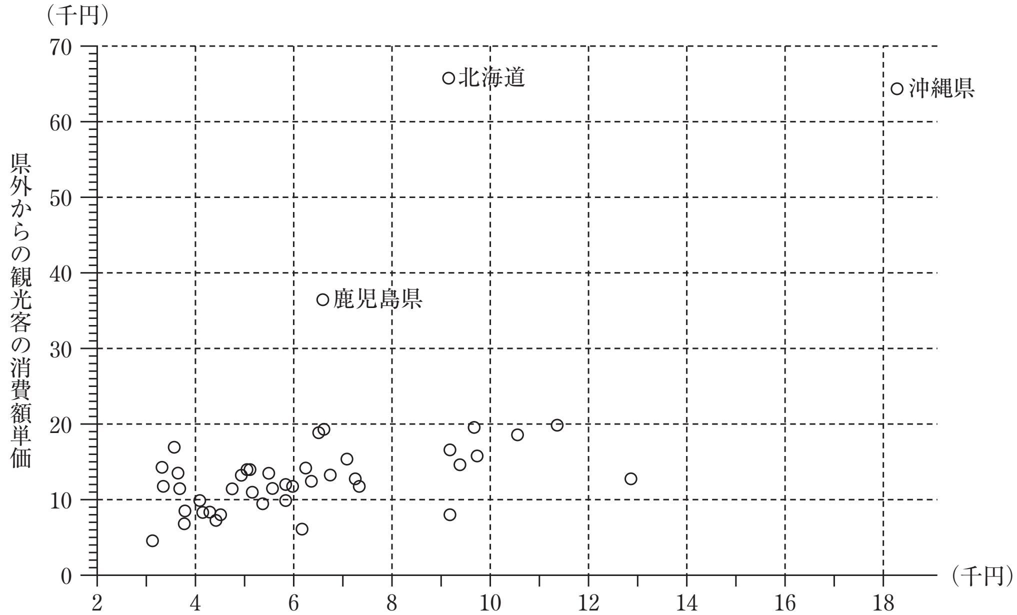


図 3

(3は次ページに続く。)

(4) 図2，図3から読み取れる事柄として正しいものを，次の⓪～④のうちから**二つ**選べ。 ウ

⓪ 44県の半分の県では，県内からの観光客数よりも県外からの観光客数の方が多い。

① 44県の半分の県では，県内からの観光客の消費総額よりも県外からの観光客の消費総額の方が高い。

② 44県の4分の3以上の県では，県外からの観光客の消費額単価の方が県内からの観光客の消費額単価より高い。

③ 県外からの観光客の消費額単価の平均値は，北海道，鹿児島県，沖縄県を除いた41県の平均値の方が44県の平均値より小さい。

④ 北海道，鹿児島県，沖縄県を除いて考えると，県内からの観光客の消費額単価の分散よりも県外からの観光客の消費額単価の分散の方が小さい。

(3は次ページに続く。)

(5)　二人は県外からの観光客に焦点を絞って考えることにした。

花子：県外からの観光客数を増やすには，イベントなどを増やしたらいいんじゃないかな。

太郎：44 県の行祭事・イベントの開催数と県外からの観光客数を散布図にすると，図 4 のようになったよ。

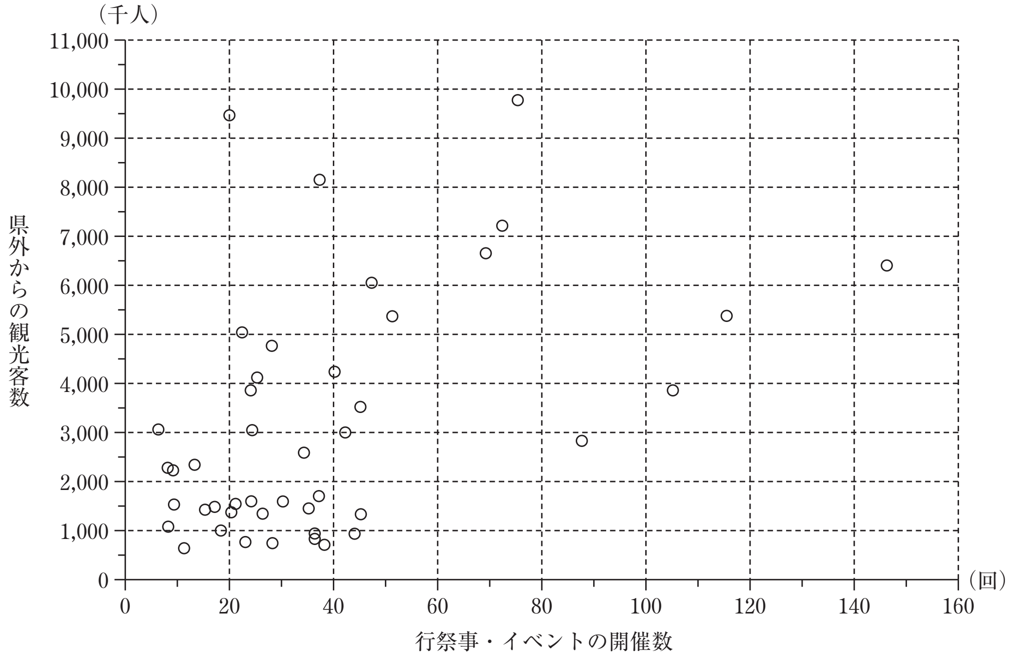


図 4

（3は次ページに続く。）

図 4 から読み取れることとして最も適切な記述を，次の⓪〜④のうちから一つ選べ。 エ

⓪ 44 県の行祭事・イベント開催数の中央値は，その平均値よりも大きい。

① 行祭事・イベントを多く開催し過ぎると，県外からの観光客数は減ってしまう傾向がある。

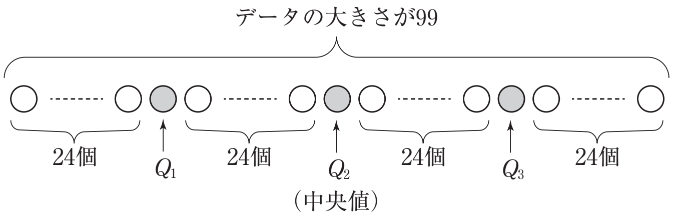
② 県外からの観光客数を増やすには行祭事・イベントの開催数を増やせばよい。

③ 行祭事・イベントの開催数が最も多い県では，行祭事・イベントの開催一回当たりの県外からの観光客数は 6,000 千人を超えている。

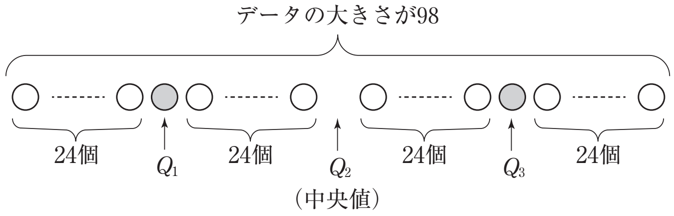
④ 県外からの観光客数が多い県ほど，行祭事・イベントを多く開催している傾向がある。

第４講 データの分析

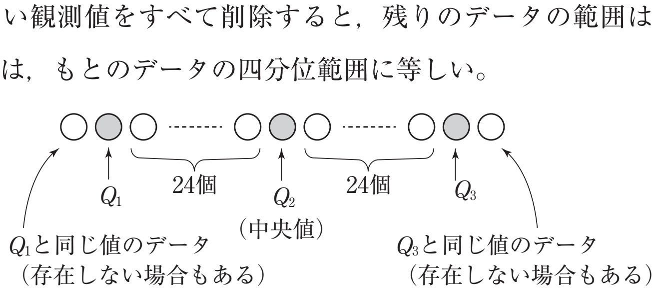
1 (1) データの大きさが 99 であるので，データを値の小さい順に並べたとき，50 番目の値が第 2 四分位数 Q_2 (中央値)である。また，1 番目から 49 番目までのデータの中央値，すなわち 25 番目の値が第 1 四分位数 Q_1 であり，51 番目から 99 番目までのデータの中央値，すなわち 75 番目の値が第 3 四分位数 Q_3 である。



③：最大値に等しい観測値を 1 個削除した残り 98 個のデータについて，第 2 四分位数 Q_2 (中央値)は 49 番目と 50 番目の値の平均値である。また，1 番目から 49 番目までのデータの中央値，すなわち 25 番目の値が第 1 四分位数 Q_1 であり，50 番目から 98 番目までのデータの中央値，すなわち 74 番目の値が第 3 四分位数 Q_3 である。よって，第 1 四分位数 Q_1 は変わらない。



⑤：範囲は，データの最大値と最小値の差である。Q_1 より小さい観測値と，Q_3 より大きい観測値をすべて削除すると，残りのデータの範囲は Q_3-Q_1 である。これは，もとのデータの四分位範囲に等しい。



以上より，どのようなデータでも成り立つものは③と⑤である。

(2) (Ⅰ)：P10 の四分位範囲は 1 より大きいので誤り。

(Ⅱ)：P10 の中央値より P11 の中央値が小さいので誤り。

(Ⅲ)：P1 の最大値は 79.5 より小さく，P47 の最小値は 81.0 より大きい。よって

正しい。

以上より，(Ⅰ)，(Ⅱ)，(Ⅲ)の正誤の組合せとして正しいものは⑥

(3) 最小値が 79.5 歳以上 80.0 歳未満で，最大値が 81.5 歳以上 82.0 歳未満である箱ひげ図は④または⑤

データを値の小さい順に並べたとき，データの大きさが 20 より，中央値は 10 番目の値と 11 番目の値の平均であり，第 1 四分位数は 5 番目の値と 6 番目の値の平均である。図 2 より，6 番目の値は 80.0 歳以上 80.5 歳未満の階級に属するので，図 2 のヒストグラムに対応する箱ひげ図は④

(4) 図 3 の 5 本の直線はそれぞれ，男女の平均寿命の差が 5.5，6.0，6.5，7.0，7.5 の点の集合である。

図 3 より，差が 7.0 から 7.5 の間に 3 個のデータがあり，ヒストグラムで該当するのは③

2 (1) 変量 x の 2 つのデータについて，

平均値 $\overline{x}=\dfrac{1+2}{2}=\dfrac{3}{2}=1.50$ (⑧)

分散 $S_x{}^2=\dfrac{(-0.5)^2+(0.5)^2}{2}=(0.5)^2$

標準偏差 $S_x=\sqrt{S_x{}^2}=\sqrt{(0.5)^2}=0.50$ (⑥)

変量 y の 2 つのデータについても，平均値 $\overline{y}=1.50$，標準偏差 $S_y=0.50$ である。

また，x と y の共分散は $S_{xy}=\dfrac{(-0.5)\times0.5+0.5\times(-0.5)}{2}=-0.25$

よって，x と y の相関係数は $r=\dfrac{S_{xy}}{S_xS_y}=\dfrac{-0.25}{0.5\times0.5}=-1$ (①)

(2) $(x,\ y)=(1,\ 2),\ (2,\ 2)$ のとき，$\overline{y}=2$，$S_y=\sqrt{\dfrac{0^2+0^2}{2}}=0$ より，

変量 y の標準偏差の値が 0 であるから(③)，相関係数 $r=\dfrac{S_{xy}}{S_xS_y}$ の計算において分母が 0 になり，相関係数が計算できない。

(3) ②：$(x,\ y)=(1,\ 1),\ (2,\ 2),\ (2,\ 2),\ (2,\ 2)$ のとき，相関係数の値は 1.00 となる。よって，②は誤り

(4) 相関係数の値はあくまで散布図の点が，直線に沿って分布する(④)程度を表している。

値の組の個数が 2 のとき，相関係数の値が 1.00，−1.00，計算できない，の 3 通りに限られるのは，平面上の異なる 2 点は必ずある直線上にある(③)からである。

散布図において，上記の異なる 2 点を通る直線の傾きが正であれば，相関係数の値は 1.00 であり，直線の傾きが負であれば，相関係数の値は -1.00 である。また，2 点が同一か，2 点を通る直線が座標軸と平行になる場合は，相関係数は計算できない。

3 (1) 観光客数 (x) と消費総額 (y) の間に強い正の相関があるので，相関係数に最も近い値は選択肢④

(2) 消費額単価＝$\dfrac{\text{消費総額}}{\text{観光客数}}$ つまり $\dfrac{y}{x}$ である。よって，各県を表す点と原点を通る直線の傾きが最も大きい点を選べばよい。

(3) (2)の考え方を用いると，消費額単価が最も高い県を表す点は⑧である。

(4) ⓪：図 2 の上の箱ひげ図では，同一の県における県内と県外の比較ができないため，この結論には至らない。よって誤り。

①：⓪と同様，図 2 の下の箱ひげ図からは，この結論には至らない。よって，誤り。

②：図 3 において，県内と県外の観光客の消費額単価が等しい点を結んだ直線を考えると，それより上側の領域，つまり県外の方が県内より多い県が 44 県中 4 分の 3 以上ある。よって正しい。

③：図 3 において，北海道，鹿児島県，沖縄県の 3 県は，県外からの観光客の消費額単価が高い方から数えて上位 3 県であるから，この 3 県を除いた 41 県の平均値は，この 3 県を含む 44 県の平均値より小さい。よって正しい。

④：図 3 において，北海道，鹿児島県，沖縄県の 3 県を除いた 41 県の分布は，横軸方向の広がりに比べて縦軸方向の広がりの方が大きい。したがって，県内の方よりも県外の方が消費額単価の分散が大きいため，誤り。

(5) ⓪：必ずしも正しくない。

①：負の相関は見られない。

②：正の相関は見られるが，因果関係があるとはいえない。

③：行祭事・イベントの開催数が最も多い県について，約 145 回開催して約 6,500 千人の観光客数であり，1 回当たりの観光客数が 6,000 千人を超えてはいない。

④：正の相関が見られるので，正しい。